Computers in Biology and Medicine 182 (2024) 109128

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

FvFold: A model to predict antibody Fv structure using protein language model with residual network and Rosetta minimization

Pasang Sherpa [a], Kil To Chong [a,b,*], Hilal Tayara [c,**]

[a] Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea
[b] Advanced Electronics and Information Research Center, Jeonbuk National University, Jeonju, 54896, South Korea
[c] School of International Engineering and Science, Jeonbuk National University, Jeonju, 54896, South Korea

ARTICLE INFO

Dataset link: https://zenodo.org/records/10791148

Keywords:
Deep learning
Antibody structure prediction
mAb
IgG
Paratope
CDR
Antibody engineering
Language model
Embedding
Rosetta

ABSTRACT

The immune system depends on antibodies (Abs) to recognize and attach to a wide range of antigens, playing a pivotal role in immunity. The precise prediction of the variable fragment (Fv) region of antibodies is vital for the progress of therapeutic and commercial applications, particularly in the treatment of diseases such as cancer. Although deep learning models exist for accurate antibody structure prediction, challenges persist, particularly in modeling complementarity-determining regions (CDRs) and the overall antibody Fv structures. Introducing the FvFold model, a deep learning approach harnessing the capabilities of the ProtT5-XL-UniRef50 protein language model which is capable of predicting accurate antibody Fv structure. Through evaluations on various benchmarks, our model outperforms existing models, demonstrating superior accuracy by achieving lower Root Mean Square Deviation (RMSD) in almost all loops and Orientational Coordinate Distance (OCD) values in the RosettaAntibody benchmark, Therapeutic benchmark and IgFold benchmark compared to the previous top-performing model.

1. Introduction

Antibodies are important immune molecules that have significant commercial and therapeutic value due to their capability to bind to a diverse array of antigens. Produced by the immune system, antibodies, also known as immunoglobulins (Ig), are Y-shaped proteins that play a vital role in combating foreign substances, or antigens, entering the body. These proteins are distributed throughout various parts of the body. B cells produce antibodies when they encounter an antigen, which causes them to divide and release millions of antibodies into the bloodstream and lymph system. Monoclonal antibodies [1], created in a lab, mimic the immune system's ability to fight pathogens and are used in immunotherapy. Antibody drugs have become increasingly important in modern medicine [2], with approximately 20% of new FDA-approved drugs being antibodies. The Antibody Society has reported that as of July 1, 2022, there were 165 therapeutic antibody drugs approved or in regulatory reviews globally. These drugs are used to treat a variety of diseases, including cancer, immune, infectious, blood system, nervous system, genetic diseases, and more [2]. Overall, antibody drugs are a significant and growing part of the biomedicine field [3]. The creation of therapeutic antibody medications is very expensive and time-consuming. Obtaining antibody structures through experimental techniques such as X-ray crystallography and nuclear magnetic resonance is a time-consuming, laborious, and expensive process. However, computational methods for predicting antibody structures offer a cost-effective and efficient alternative. Computational methods enable researchers to obtain antibody structures easily that may be difficult or time taking through experimental techniques.

Recent advancements in deep learning have significantly enhanced the accuracy of general protein structure prediction. AlphaFold-Multimer [4], a notable example, achieves impressive results without specific optimization for the antibody sub-domain. Deep learning has revolutionized protein structure prediction, making it more accessible and cost-effective than traditional experimental methods. In the domain of antibody protein structure prediction, various deep-learning models have emerged, including IgFold [5], DeepAb [6], AlphaFold-Multimer [4], and ABlooper [7]. IgFold employs a rapid end-to-end deep learning model, integrating pre-trained BERT language models. AlphaFold-Multimer utilizes a transformer-based architecture to predict multichain protein complexes. DeepAb employs a deep residual convolutional network along with a Rosetta-based protocol for precise

* Corresponding author at: Department of Electronics and Information Engineering, Jeonbuk National University, Jeonju, 54896, South Korea.
** Corresponding author.
E-mail addresses: kitchoong@jbnu.ac.kr (K.T. Chong), hilaltayara@jbnu.ac.kr (H. Tayara).

https://doi.org/10.1016/j.compbiomed.2024.109128
Received 24 May 2024; Received in revised form 22 August 2024; Accepted 5 September 2024
Available online 12 September 2024

Complementarity-determining regions (CDRs) loop modeling. On the other hand, ABlooper predicts CDR loop structures using an end-to-end adapted graph neural network architecture.

In this paper, we have developed the FvFold model with the aim of enhancing the accuracy of antibody Fv structure prediction through the utilization of a protein language model [8]. Our approach involves extracting per-residue embeddings for heavy and light chain sequences using the ProtTrans protein language model, specifically ProtT5-XL-UniRef50 [9]. These embeddings are fed into a deep residual network to predict inter-residue distances and orientations. After that, the predicted structure is processed by the Rosetta protocol to enhance the prediction of three-dimensional antibody structures from sequence data. The experimental verification of our model against various benchmarks, including IgFold [5] benchmark set, RosettaAntibody benchmark set which has previously been used in previous methods [10–12], and therapeutic benchmarks dataset which was previously assembled to study antibody developability [13], provides compelling evidence that our model excels in predicting overall antibody Fv structures.

2. Materials and methods

2.1. Data collection and preprocessing

The training dataset was download from a curated database of all antibody structures known as Structural Antibody Database (SAbDab) [14,15]. SAbDab stands as a centralized repository of antibody structures sourced from the Protein Data Bank (PDB), offering standardized annotations encompassing experimental details, affinity data, and sequence information. This invaluable resource caters to researchers engaged in the study of antibodies. In total, the database contained approximately 7,689 antibody PDB entries. However, we specifically extracted a non-redundant paired set of antibodies from SAbDab, focusing on structures annotated using the Chothia numbering scheme. We filtered structures with less than 4-Åresolution, complex type both (bound and unbound) and applied a 99% sequence identity threshold to remove redundant sequences. Following the download of the complete antibody PDB file from the database, consisting of 2,438 entries, we conducted a filtering process. Specifically, we extracted the Fv (Fragment Variable) by eliminating excess residues, refining the dataset for further analysis or application as a template. The filtration of the PDB dataset involved two key criteria. Firstly, PDB entries containing both heavy and light chains together were excluded. Secondly, PDB files with missing chains were also removed. This filtering process aimed to ensure a refined and reliable dataset for subsequent analyses. In cases where a PDB file contained multiple chains for the same structure, only the first chain was utilized. Additionally, to further refine the dataset, all structures present in both the Therapeutic Benchmark dataset and the Rosetta Antibody Benchmark dataset were systematically eliminated. This comprehensive filtering strategy enhances the specificity and relevance of the selected antibody structures for subsequent analyses or applications. Furthermore, we acquired a FASTA file from the Protein Data Bank (PDB) [16] using the corresponding PDB identifiers. Following this, we truncated the obtained sequences to isolate the Fragment Variable (Fv) region. This process resulted in the creation of 2,253 FV structures along with a corresponding FASTA file. These structures and sequence data were utilized for the training of our model. During the training phase, 90% of the dataset was allocated for training the model, with the remaining 10% reserved for validation purposes. Testing was conducted on a separate benchmark, specifically the Rosetta and Therapeutic datasets, ensuring that there were no shared structures with the training or validation sets. To enhance model robustness and performance, we implemented five training iterations. In each iteration, both the training and validation sets were shuffled, resulting in the creation of five distinct sets of weights. These weights were saved individually and later aggregated through an ensemble approach. This ensemble of weights was utilized to predict the Fragment Variable (Fv) structure of antibodies, contributing to a more reliable and accurate model outcome.

2.2. Benchmark datasets

To evaluate the effectiveness of our methodology, we employed two separate sets of evaluations, ensuring their complete independence. The first set involved the use of the RosettaAntibody [17] benchmark, which comprised 49 structures [12]. The second set consisted of a distinct compilation comprising 56 clinical-stage antibody therapeutics, analyzed through solved crystal structures. The independence of these evaluation sets adds robustness to the assessment of the model's performance across different benchmarks and real-world clinical scenarios. Nevertheless, certain targets were excluded from the evaluation for various reasons. Within the RosettaAntibody benchmark, targets PDB 1X9Q and 3IFL were excluded due to the absence of structures predicted by any method. Additionally, PDB 3MLR was excluded because it contained a long L3 loop that was not modeled by some methods. In the therapeutic benchmark, certain targets were excluded from the evaluation for various reasons. Specifically, targets PDB 4D9Q and 4K3J were removed as no structures were predicted by some method. Additionally, PDB 4O02 and 5VVK were excluded from RepertoireBuilder and ABodyBuilder predictions. Furthermore, targets PDB 3B2U, 3C08, 3HMW, 3S34, 4EDW, 3EO9, and 3GIZ were omitted due to missing one or more Complementarity-Determining Region (CDR) loops or being common between both benchmarks. In total, metrics were reported for 91 targets, with 46 originating from the RosettaAntibody benchmark and 45 from the therapeutic benchmark. This approach ensures a fair and consistent comparison across all methods, providing the model's predictive capabilities for the given set of targets. Throughout this evaluation, we employed the Chothia Complementarity-Determining Region (CDR) loop definitions to measure Root Mean Square Deviation (RMSD) [18]. Our approach was systematically compared against five alternative methods, namely RosettaAntibody-G [17], RepertoireBuilder [19], ABodyBuilder [20], DeepAb [6], and IgFold [5].

In preparation for the evaluation against the IgFold benchmark, we conducted additional training with a specific focus on optimizing performance for this benchmark. To ensure a fair and unbiased comparison, we systematically excluded all 197 PDB entries associated with the IgFold benchmark from the initial training set. This exclusion resulted in a refined dataset comprising 2,213 entries. Then we trained the model and proceeded to evaluate the model's predictive accuracy by employing the IgFold benchmark dataset. The utilization of multiple benchmark datasets and the inclusion of various antibody-specific methodologies allowed us to thoroughly evaluate and benchmark the effectiveness and efficiency of our method. This comparison provided valuable insights into the strengths and weaknesses of our approach in comparison to established antibody analysis models.

2.3. Feature preparation (ProtTrans)

In our deep learning FvFold model, we utilized ProtTrans (ProtT5-XL-UniRef50) [9] model to extract per residue embeddings, ProtTrans is providing state-of-the-art pre-trained models for proteins. It was pretrained on protein sequences using a masked language modeling (MLM) objective, leveraging a dataset comprising billions of protein sequences [21,22]. ProtTrans demonstrates outstanding proficiency in processing protein sequences. Through training on an extensive dataset of protein sequences, it acquires a deep understanding of patterns and associations among diverse amino acids, capturing both their functional and structural attributes. Since complex model requires a lot of computing power, we started the ProtT5-XL-UniRef50 [9] feature extraction process for the FASTA sequences in advance. The extracted features were stored in a separate file. This approach not only prevented unnecessary parallel processing during model training but also helped mitigate the issue of excessive GPU consumption. Consequently, our model's overall computational burden was reduced and was significantly optimized feature extraction process for the FASTA sequences in advance. The ProtT5-XL-UniRef50 extraction process tends to run

in parallel with our model during model predictions. This parallel execution, although computationally intensive, ensures that our model is capable of effectively handling a diverse range of sequence inputs during real-time predictions.

2.4. Feature preparation (ResNet)

In the preprocessing pipeline for input data before passing it to the ResNet1D [23,24] model, a series of transformations are applied to convert amino acid sequences from their single-letter codes into a format suitable for deep learning analysis. This involves several steps to ensure that the sequence data is appropriately represented for processing. Initially, the amino acid sequence is represented using a dictionary, which maps each amino acid's single-letter code to a corresponding numerical index. This mapping serves as a bridge between the symbolic amino acid characters and their corresponding numerical values, allowing for a more structured representation. Following this, the next step involves the conversion of the numerical indices into one-hot encoded vectors. One-hot encoding is a technique that transforms categorical data into a binary vector representation. This conversion ensures that the ResNet1D model can effectively process the sequence data, as deep learning models typically operate more efficiently on numerical inputs. Once the one-hot encoding has been applied to the entire sequence, the resulting binary vectors, now representing the amino acids in a structured numerical format, are then fed into the ResNet1D architecture for further analysis. ResNet1D, a variant of the Residual Network (ResNet) model designed specifically for one-dimensional sequence data, can effectively capture complex patterns and dependencies within the amino acid sequences.

2.5. Models overview

In this work, we present the FvFold model, showcasing its advancements over previous models, demonstrating its capabilities to predict Fragment Variable (Fv) structure of antibody accurately.

In Fig. 1, the process starts with the retrieval of Chothia Numbered PDB(Protein Data Bank) files from the SabDab(Structural Antibody Database) database [14], representing structural information of antibodies. Additionally, FASTA files are acquired from the RCSB database [25], providing the primary sequences of the heavy and light chains corresponding to specific PDB IDs. The obtained PDB files are preprocessed, wherein irrelevant structural details are eliminated, leading to the isolation of Fv structures. As input, the deep learning model takes the FASTA files, containing the sequences of the heavy and light chains, enabling the extraction of significant structural insights from the antibody Fv regions [26]. Our model draws inspiration from the DeepAb [6] model, with similarities in the deep learning approach. However, a notable distinction lies in the training dataset size, as DeepAb utilized 1692 dataset, whereas our model incorporates a more extensive dataset of 2269. Additionally, we have implemented a ProtTrans, a protein language model to extract per-residue embeddings, a feature not present in the original DeepAb model. This addition contributes to the improvement of the DeepAb deep learning model. Essentially, our model represents is an enhanced version of the DeepAb model, aiming to predict the Fv structure of antibodies more accurately. The model architecture features two parallel branches: 1D ResNet and Protein Language model. The input data for our 1D ResNet model consists of amino acid sequences [24]. A one-hot encoding process [27] is applied before forwarding the amino acid sequence to the 1D ResNet model. On the other hand, the Protein Transformer (ProtT5-XL-UniRef50) [9] directly takes the original amino acid sequences of the heavy and light chains as input and utilizes self-attention mechanisms to capture long-range dependencies between amino acids, allowing it to learn complex spatial relationships within the antibody. The embeddings generated by both the 1D ResNet and the Protein Transformer are concatenated into a single feature representation. This concatenation merges the learned features from both models, effectively capturing complementary information. The resulting concatenated embeddings are then transformed from 1D to 2D, creating a 2D representation of the antibody. This transformation retains the sequential information while incorporating spatial relationships between different regions of the heavy and light chains. The 2D representation of the antibody is processed by a 2D ResNet model, specifically designed to handle 2D data like the transformed embeddings. The 2D ResNet employs convolutional layers to extract spatial features and patterns from the 2D data. Additionally, a Criss-Cross Attention layer is applied to the 2D ResNet to enable the model to focus on inter-region dependencies and interactions within the antibody structure effectively. The 2D ResNet with Criss-Cross Attention predicts six output values: inter-residue distances between three pairs of atoms (C_α–C_α, C_β–C_β, N–O), as well as the set of inter-residue dihedrals (ω: C_α–C_β–C_β–C_α, θ: N–C_α–C_β–C_β) and planar angles (φ: C_α–C_β–C_β) [28]. These predicted values encode critical geometric information about the positions and angles between specific atoms in the 3D structure of the antibody [29]. The predicted inter-residue distances, dihedrals, and planar angles serve as constraints in a Rosetta-based protocol for 3D structure prediction. It utilizes energy functions and optimization algorithms to model the 3D structure of proteins based on the given constraints and information.

2.6. Multi-dimensional scaling

Similar to previous approaches for predicting general protein structure [28–30], we used predictions obtained from our network and formulate real-value matrices representing the outputs for d_{C_β}, ω, θ, and φ by determining the midpoint value within each modal probability bin for every pair of residues. Using these real-valued distances and orientations, we construct an initial matrix indicating the distances between backbone atoms (N, C_α, and C). In cases where residue pairs are projected to possess d_{C_β} values exceeding 18 Å, we estimate the inter-atom distances using the Floyd-Warshall shortest path algorithm [31]. This algorithm facilitates the calculation of distances between atoms based on the connectivity of the protein structure. Utilizing the derived distance matrix, we employ Multi-dimensional Scaling (MDS) to generate an initial collection of three-dimensional coordinates [32]. The MDS technique translates the calculated distances into spatial coordinates that represent the protein's structural arrangement.

2.7. Energy minimization refinement

The refinement process starts with initial structures obtained through Multi-Dimensional Scaling (MDS), and these structures are subsequently subjected to constrained energy minimization within the Rosetta framework. For each pair of residues in the protein, the projected distributions for each output are transformed into energy potentials through the negation of raw model logits (bypassing the softmax activation) and division by the square of the predicted distance between C_α atoms (d_{C_α}). The discrete potentials are then transformed into continuous functions using Rosetta's built-in spline function. Certain potentials are omitted from consideration: residue pairs with predicted d_{C_α} distances greater than 18 Å are excluded, along with those presenting a modal bin probability lower than 10%. For potentials related to the distance between nitrogen and oxygen atoms ($d_{\text{N-O}}$), predictions exceeding 5 Å or model bin probabilities below 30% are discarded, establishing a local potential governing backbone hydrogen-bonding interactions. The potentials that meet the selection criteria are employed as inter-residue constraints on the MDS-generated structure within the Rosetta framework [33]. This energy minimization refinement is inspired from previous paper DeepAb [6]. The modeling process in Rosetta is initiated using a coarse-grained representation, wherein side-chain atoms are condensed into a singular artificial atom referred to as the centroid. Optimization of the centroid model is achieved through gradient-based

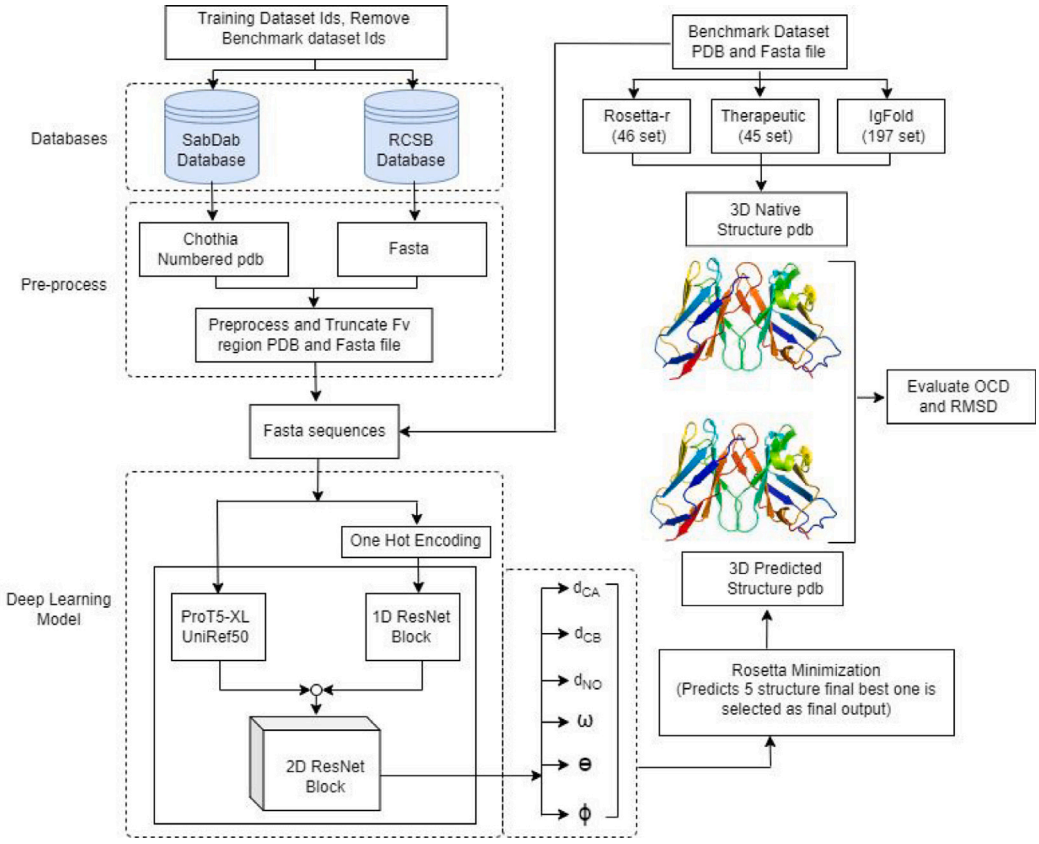


Fig. 1. A schematic representation illustrates deep learning-based approach for predicting the Fragment Variable (Fv) structure of antibodies using input sequences from the heavy and light chains in a FASTA file. The deep learning model receives these sequences as input and predicts three inter-residue distances, two inter-residue dihedrals, and one planar angle. The resulting output is fed into Rosetta protocols to generate 3D structure of the antibody.

energy minimization utilizing the MinMover algorithm and the L-BFGS optimization technique [33,34]. The energy function for the centroid includes additional scoring terms alongside the learned constraints, encompassing van der Waals interactions (clashes), centroid-based hydrogen bonding (cen_hb), and evaluations of backbone torsion angles (rama and omega). Following centroid optimization, side-chain atoms are reintroduced, and the structure undergoes relaxation to alleviate steric clashes via the FastRelax procedure. Subsequently, a repetition of the gradient-based energy minimization step is executed in the full-atom representation to yield the final model. To ensure comprehensive conformational sampling, this refinement protocol is reiterated to generate 50 decoy models. Among these decoys, the structure with the lowest energy value is selected as the final prediction. Notably, the only non-deterministic aspect of the protocol occurs during the relaxation step, fostering a high level of convergence among the decoy models. In practical terms, a modest number of 5–10 decoy models are generally sufficient for most applications.

2.8. Evaluation metrics (RMSD and OCD)

2.8.1. RMSD (Root Mean Square Deviation)

In this study, we employed Root Mean Square Deviation (RMSD) analysis to assess the structural differences between the antibody regions of interest and a native PDB structure. RMSD is a widely accepted metric in structural biology for quantifying the similarity or deviation between two protein structures. In our study, the analyzed regions encompassed the Heavy Framework Region (HFr), H1, H2, H3, Light Framework Region (LFr), L1, L2, and L3, similar to the regions studied in the DeepAb [6] and IgFold [5] papers. The RMSD values were computed by aligning the corresponding regions of our predicted antibody structure to the native PDB structure and measuring the average root mean square displacement of the atomic positions. The resulting RMSD values provide insights into the structural variations between our antibody and the native structure. These values are presented in evaluation tables.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

N: The number of data points or observations in the dataset.

x_i: Individual observed values in the dataset.

$\hat{x}_i$: Corresponding predicted or modeled values for each observed value. The formula calculates the average squared difference between each observed and predicted value, sums these squared differences, divides by the number of data points, and takes the square root to get the RMSD. It provides a quantitative measure of how well a model or prediction fits the actual data. A lower value indicates a closer structural similarity between the predicted antibody structure and the reference PDB structure. On the other hand, a higher RMSD value suggests a greater structural deviation, indicating differences between the predicted and actual structures.

2.8.2. OCD (Orientational Coordinate Distance)

The use of Orientation Coordinate Deviation (OCD) values, as introduced in the DeepAb [6] and IgFold [5] papers, is another metric employed in our study. This metric is determined by summing the deviations in orientation for four coordinate parameters from their native values, which include the packing angle, interdomain distance, heavy-opening angle, and light-opening angle. The resulting sum is subsequently divided by the corresponding standard deviation for each coordinate. This approach facilitates an evaluation of the alignment between the predicted orientation and the native configuration of the

Table 1
Therapeutic benchmark (45 dataset).

Method	OCD	HFr	H1	H2	H3	LFr	L1	L2	L3
RosettaAntibody-G	5.43	0.63	1.42	1.05	3.77	0.55	0.89	0.83	1.48
RepertoireBuilder	4.37	0.62	0.91	0.96	3.13	0.47	0.71	0.52	1.08
ABodyBuilder	4.37	0.49	1.05	1.02	3.00	0.45	1.04	0.50	1.35
IgFold	3.69	0.47	0.89	0.81	2.41	0.48	0.79	0.60	1.04
DeepAb	3.52	0.40	0.77	0.68	2.52	0.37	0.60	0.42	1.02
FvFold	3.09	0.39	0.75	0.67	2.33	0.35	0.67	0.38	0.96

heavy and light chains. It is also an important factor in predicting the conformation of the complementarity-determining region (CDR) H3 loop, which is a crucial component of the antibody's binding site. The H3 loop's position between the heavy and light chains makes it dependent on the chain orientation and multiple adjacent loops [12].

The Orientational Coordinate Distance (OCD) is calculated using the following formula:

$$\mathrm{OCD} = \sum_{i=\alpha,\delta_{\mathrm{ID}},\theta_{(L)},\theta_H} \frac{(x_{i,A} - x_{i,B})}{\sigma_{i,\mathrm{dB}}}$$

In this formula:

OCD represents the Orientational Coordinate Distance, which is a dimensionless metric. i is a variable that takes on four specific values: α, δ_{ID}, θ_{L}, and θ_{H} (packing angle, interdomain distance, heavy-opening angle, and light-opening angle).

$x_{i,A}$ is the value of the LHOC metric i for structure A.

$x_{i,B}$ is the value of the LHOC metric i for structure B.

$\sigma_{i,\mathrm{dB}}$ represents the standard deviation of the Gaussian distribution that best fits the database distribution of the LHOC metric I. This formula calculates the OCD by summing up the deviations of the four orientation coordinates (packing angle, interdomain distance, heavy-opening angle, and light-opening angle) for structures A and B, divided by the respective standard deviations for each of these coordinates. The outcome quantifies the accuracy of the heavy-light chain orientation between the two structures.

3. Results

In our study, we conducted performance evaluation of our method using three benchmark datasets: the RosettaAntibody benchmark, Therapeutic benchmark and IgFold benchmark. To provide insight of our approach, we compared it against five alternative methods that are specifically designed for antibody modeling. These methods include RosettaAntibody-G [17], RepertoireBuilder [19], ABodyBuilder [20], DeepAb [6], and IgFold [5]. Similar to the DeepAB method, we systematically excluded all 46 PDB entries from the RosettaAntibody Benchmark and 45 PDB entries from the Therapeutic Benchmark from our initial training dataset. We evaluated various antibody structure prediction methods using the Orientational Coordinate Distance (OCD) and Root Mean Square Deviation (RMSD) metrics to assess their accuracy in predicting 3D antibody structures compared to native conformations.

3.1. FvFold outperforms previous methods in therapeutic benchmark

FvFold demonstrated exceptional performance on the Therapeutic Benchmark Dataset, from Table 1 we can see that our method achieved the lowest OCD value of 3.09, indicating superior accuracy in predicting antibody orientations compared to native conformations. FvFold outperformed other models in capturing heavy chain segments, with the lowest RMSD values for HFr (0.39 Å), H1 (0.75 Å), H2 (0.67 Å), and H3 (2.33 Å). For light chain regions, FvFold exhibited high accuracy, achieving low RMSD values for LFr (0.35 Å), L2 (0.38 Å), and L3 (0.96 Å). In specific segments, DeepAb excelled, achieving the lowest RMSD value of 0.60 Åfor the L1 region.

Table 1: FvFold surpasses other methods in Orientational Coordinate Distance (OCD) and exhibits lower Root Mean Square Deviation (RMSD) values across almost all Complementarity-Determining Region (CDR) loops on the RosettaAntibody Benchmark Datasets.

Table 2
Rosetta antibody benchmark (46 dataset).

Method	OCD	HFr	H1	H2	H3	LFr	L1	L2	L3
RepertoireBuilder	5.25	0.58	0.88	1.00	2.88	0.51	0.63	0.53	1.03
RosettaAntibody-G	5.23	0.57	1.23	1.14	3.49	0.57	0.75	0.85	1.06
ABodyBuilder	4.71	0.5	1.00	0.87	2.89	0.49	0.71	0.51	1.09
IgFold	4.18	0.49	0.80	0.83	2.65	0.51	0.67	0.60	0.97
DeepAb	3.70	0.43	0.72	0.85	2.31	0.42	0.55	0.46	0.83
FvFold	3.23	0.41	0.71	0.76	2.39	0.40	0.56	0.42	0.81

Table 3
IgFold benchmark (197 dataset).

Method	OCD	HFr	H1	H2	H3	LFr	L1	L2	L3
RepertoireBuilder	5.09	0.59	1.00	0.90	4.15	0.49	0.81	0.57	1.32
ABlooper	4.42	0.53	0.98	0.83	3.54	0.51	0.92	0.67	1.32
AlphaFold-Multimer	4.18	0.69	0.95	0.74	3.56	0.66	0.84	0.51	1.59
IgFold	3.82	0.48	0.85	0.76	3.27	0.46	0.76	0.46	1.3
DeepAb	3.60	0.43	0.86	0.72	3.57	0.41	0.75	0.48	1.16
FvFold	3.39	0.42	0.83	0.69	3.26	0.39	0.67	0.42	1.06

According to above results, Overall FvFold showcased superior accuracy and orientation alignment, making it a promising model for precise antibody structure prediction. To further confirm whether our model was really stable model we compared our model with other benchmarks datasets as well.

3.2. FvFold outperforms previous methods in RosettaAntibody benchmark

Table 2 FvFold demonstrates superior performance in Orientational Coordinate Distance (OCD) and exhibits lower RMSD values across almost all Complementarity-Determining Region (CDR) loops on the Rosetta Benchmark Datasets compared to other methods.

As shown in Table 2 for RosettaAntibody benchmark dataset, our method FvFold, exhibited exceptional performance in predicting antibody orientations, achieving the lowest OCD value of 3.23. This result signifies its superior alignment with native structures. In capturing heavy chain segments, FvFold outperformed other models, showcasing the lowest RMSD values for Heavy Framework Region (HFr) (0.41 Å), H1 (0.71 Å), and H2 (0.76 Å). For the light chain regions, FvFold excelled, attaining low RMSD values for Light Framework Region (LFr) (0.40 Å), L2 (0.42 Å), and L3 (0.81 Å). In specific segments, DeepAb achieved the lowest RMSD values, measuring 2.31 Åfor the H3 region and (0.55 Å) for the L1 region.

Overall, FvFold demonstrated the capability to predict loops more effectively, as evidenced by this benchmark, showcasing its accuracy in predicting the overall Fragment Variable (Fv) structure of antibodies. Although in this benchmark dataset our model was not able to get top CDRH3 value but it is quite close to DeepAb results.

3.3. FvFold outperforms previous methods in IgFold benchmark

We conducted additional training for our model to show the consistency of our model on the IgFold [5] benchmark. To ensure a fair comparison of results, we systematically removed all 197 PDB (Protein Data Bank) entries associated with the IgFold benchmark from our initial training dataset, leaving us with a refined dataset of 2213 entries. Finally, we evaluated the model using the IgFold benchmark dataset, aiming for a comprehensive and unbiased assessment of its predictive capabilities in comparison to this specific benchmark.

In Table 3, FvFold surpasses other methods on IgFold Benchmark Datasets, showcasing its superior performance in this evaluation. Similarly, in the IgFold benchmark, FvFold demonstrated superior performance, achieving the lowest OCD value of 3.39 and the lowest RMSD values for all individual loops. Lower OCD and RMSD values indicate

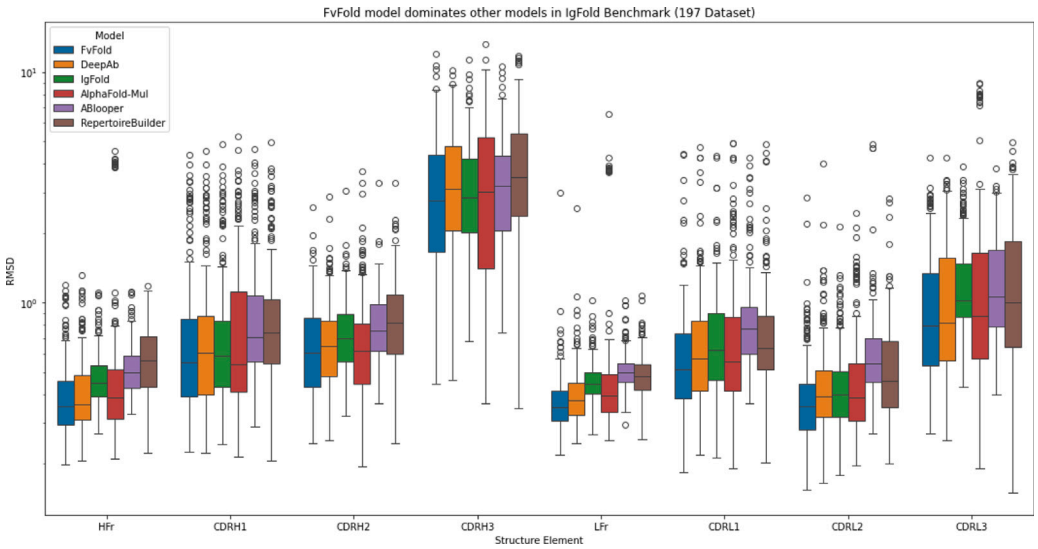
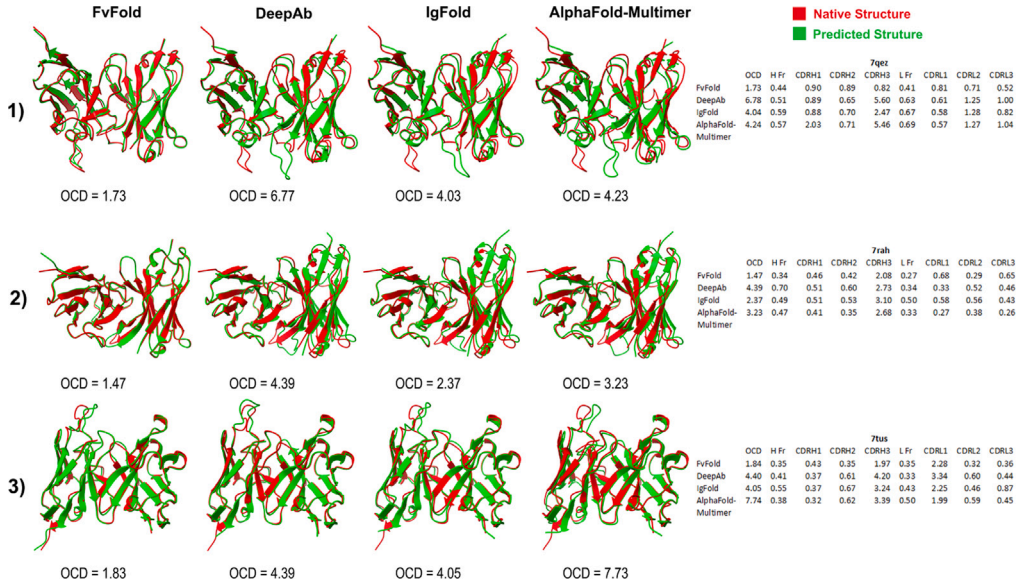


Fig. 2. Benchmark performance of RepertoireBuilder, DeepAb, ABlooper, AlphaFold-Multimer, IgFold and FvFold for paired antibody structure prediction in IgFold benchmark datasets (197 structures).



7qez

	OCD	H Fr	CDRH1	CDRH2	CDRH3	L Fr	CDRL1	CDRL2	CDRL3
FvFold	1.73	0.44	0.90	0.89	0.82	0.41	0.81	0.71	0.52
DeepAb	6.78	0.51	0.89	0.65	5.60	0.63	0.61	1.25	1.00
IgFold	4.04	0.59	0.88	0.70	2.47	0.67	0.58	1.28	0.82
AlphaFold-Multimer	4.24	0.57	2.03	0.71	5.46	0.69	0.57	1.27	1.04

7rah

	OCD	H Fr	CDRH1	CDRH2	CDRH3	L Fr	CDRL1	CDRL2	CDRL3
FvFold	1.47	0.34	0.46	0.42	2.08	0.27	0.68	0.29	0.65
DeepAb	4.39	0.70	0.51	0.60	2.73	0.34	0.33	0.52	0.46
IgFold	2.37	0.49	0.51	0.53	3.10	0.50	0.58	0.56	0.43
AlphaFold-Multimer	3.23	0.47	0.41	0.35	2.68	0.33	0.27	0.38	0.26

7tus

	OCD	H Fr	CDRH1	CDRH2	CDRH3	L Fr	CDRL1	CDRL2	CDRL3
FvFold	1.84	0.35	0.43	0.35	1.97	0.35	2.28	0.32	0.36
DeepAb	4.40	0.41	0.37	0.61	4.20	0.33	3.34	0.60	0.44
IgFold	4.05	0.55	0.37	0.67	3.24	0.43	2.25	0.46	0.87
AlphaFold-Multimer	7.74	0.38	0.32	0.62	3.39	0.50	1.99	0.59	0.45

Fig. 3. Antibody Fv structure prediction for 3 targets 1) 7qez, 2) 7rah, and 3) 7tus. Representing red as native structure and green as predicted structure by FvFold, DeepAb, IgFold, and AlphaFold-Multimer models.

better accuracy, establishing FvFold as the most accurate method overall. This reinforces FvFold's standing as a promising model for precise antibody structure prediction, characterized by superior accuracy and orientation alignment compared to other methods.

In Fig. 2 the comparison includes root-mean-squared-deviation (RMSD) values calculated over backbone heavy atoms after alignment of the respective framework residues. Box plots shows the median, interquartile range, and outliers. We can see that FvFold model outperforms other models, achieving lower RMSD values in average across all loops for paired 197 Fv antibody structures in the IgFold benchmark.

In Fig. 3, it shows a comparison of predictions from four models: FvFold, DeepAb, IgFold, and AlphaFold-Multimer. The native structures of Three antibodies (1) 7qez, (2) 7rah, and (3) 7tus are highlighted in red. Each native antibody structure is paired with predictions from different models, shown in green. The accompanying side table provides detailed values for OCD, Fr RMSD, and CDR RMSD, offering a comprehensive comparison with the native structure. From Figs. 2 and 3 and Tables 1, 2 and 3, we can see our model has been able to predict overall Fv structure of the antibody accurately obtaining lower RMSD and OCD values.

3.4. CDRH3 loop comparison

From Fig. 4 the boxplot shows that our method FvFold excelled in the Therapeutic benchmark, achieving the best performance in predicting CDR H3 loop structures with an average RMSD of 2.33 Å. Similarly, on the IgFold benchmark, our method FvFold outperformed others, attaining the top position with an average CDR H3 loop RMSD of 3.25 Å. However, it is worth noting that on the RosettaAntibody Benchmark, the DeepAb model surpassed others, demonstrating superior performance by predicting CDR H3 loop structures with an average RMSD of 2.3 Åwhereas our method stood second having average RMSD of 2.39. From Fig. 4B the predicted structure of different model have been compared with the native structure. And the diagram shows how our model is capable of predicting accurate antibody Fv structure.

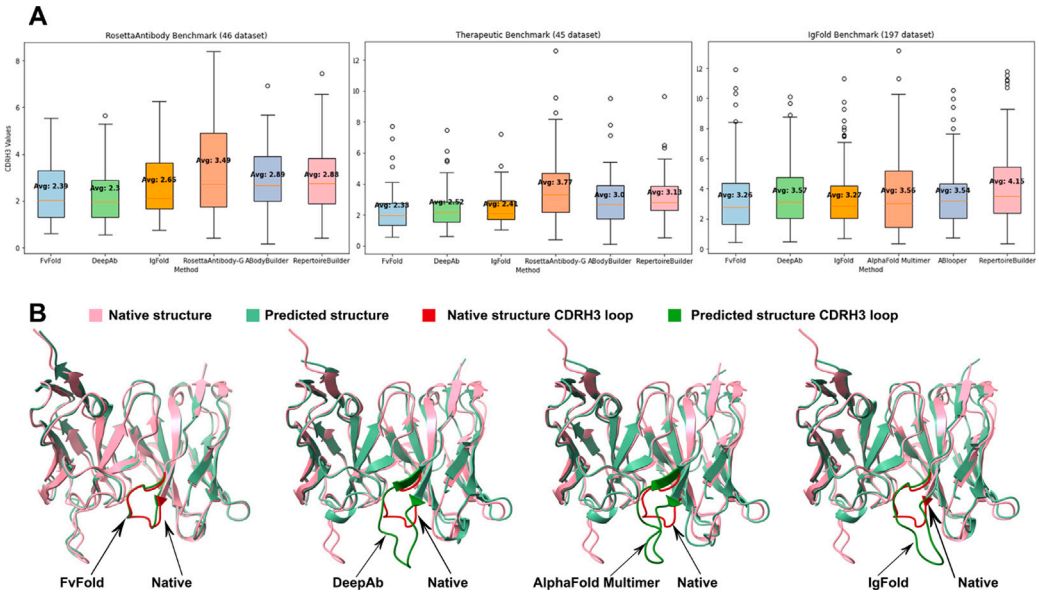


Fig. 4. (A) Comparison of CDRH3 loops predicted by different model on the RosettaAntibody benchmark, Therapeutic benchmark and IgFold benchmark (B) Comparison of native CDRH3 loop structure (red, PDB: 7qez) to predictions from FvFold(green, 0.82 ÅRMSD), DeepAb(green, 5.60 ÅRMSD), AlphaFold Multimer(green, 5.46 ÅRMSD) and IgFold(green, 2.47 ÅRMSD). We can see that our FvFold model predicts accurate structure for these cases.

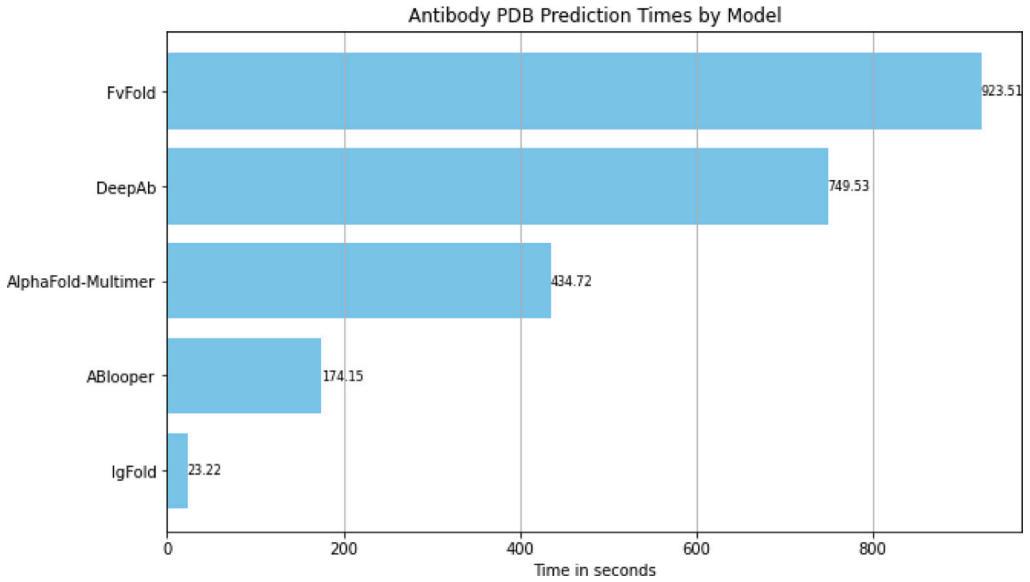


Fig. 5. The chart illustrates the varying execution times (in seconds) for different models to predict antibody 3D structure. IgFold exhibit relatively quick processing, while AlphaFold-Multimer and DeepAb take longer. FvFold takes the longest prediction time, indicating diverse computational demands across the models.

3.5. Comparative analysis of antibody structure prediction times

Among the tested methods for paired antibody structure prediction on IgFold benchmark set(197dasets), IgFold demonstrated superior speed, taking an average of 23.22 s to predict a full-atom structure from sequence. In comparison, ABlooper, the next fastest method, required nearly 3 min (174.15 s) for full-atom structure prediction. Although ABlooper rapidly predicted coordinates in an end-to-end fashion, subsequent refinement in OpenMM was necessary to correct geometric abnormalities and add side chains, contributing to its overall longer runtime. The ColabFold(26) implementation of AlphaFold-Multimer, in this evaluation, exhibited an average prediction time just over 7 min (434.7 s) for full-atom structure prediction. Notably, this implementation outperformed the original AlphaFold-Multimer, which incurred prolonged prediction times due to an expensive MSA search and repeated model compilation for each prediction. Conversely, DeepAb and FvFold, the slowest methods for paired antibody structure prediction, had an average runtime exceeding 12 min (749.53 s) and 15 min (923.51 s). DeepAb's and FvFold's deliberate design, involving the minimization of predicted inter-residue potentials in Rosetta, contributed to its inherently slower computational speed. Our model FvFold also contains protein language model for feature extraction which makes it slowest among the above methods. Additionally, an analysis of the impact of sequence length on prediction times revealed a general trend of increased runtimes for all methods with longer sequences. FvFold, DeepAb and ABlooper proved to be the most sensitive to sequence length, while AlphaFold-Multimer and IgFold exhibited more favorable scaling (see Fig. 5).

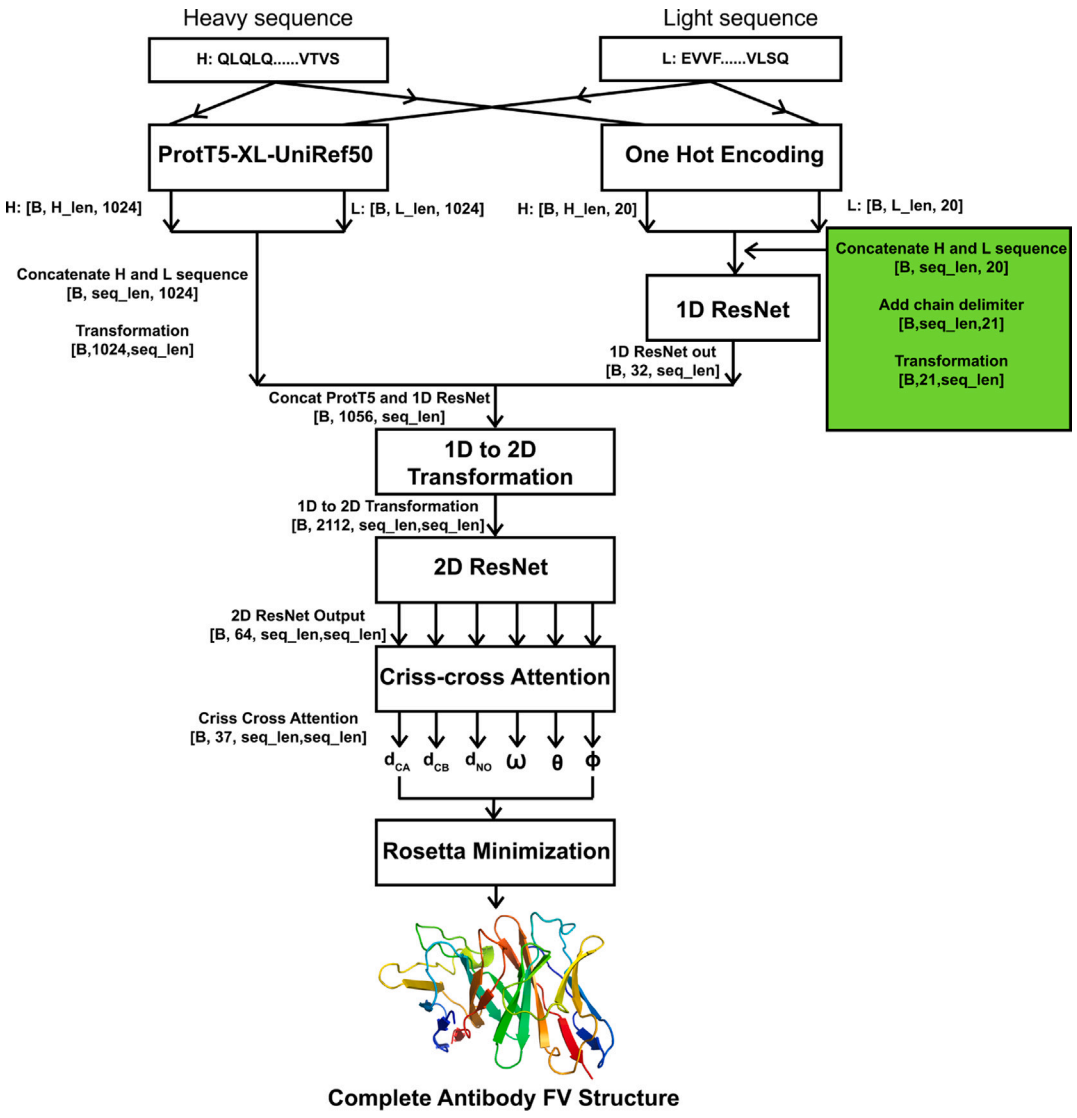


Fig. 6. FvFold Model Architecture.

3.6. Models detail

The antibody heavy and light chain sequences were input separately into the ProtT5-XL-UniRef50 model to obtain per residue embeddings. The resulting input dimensions were [B,H_len,1024] and [B,L_len,1024], where B represents the batch size, H_len is the length of the heavy chain sequence, and L_len is the length of the light chain sequence. To combine the heavy and light chain sequences, they were concatenated, resulting in a tensor with dimensions of [B, seq_len,1024], where seq_len represents the total sequence length of the combined heavy and light chains. This tensor was then transformed to [B, 1024, seq_len]. In addition, the same heavy and light chain sequences were one-hot encoded separately and concatenated again. A chain delimiter, represented by 0, was added at the last position of the heavy chain and introduced an additional delimiter channel to mark this end of heavy chain. This delimiter facilitates the segmentation of data into discrete units, helping to identify where one sequence ends and another begins. This allows for accurate data parsing and manipulation, making it easier for the model to separate the heavy and light chains. This resulted in a tensor with dimensions of [B, seq_len,21], where seq_len represents the total length of the combined heavy and light chain sequences. The tensor was then transformed to [B, 21, seq_len] (see Fig. 6).

The concatenated vectors from the ProtT5 model and the 1D ResNet model were combined, resulting in a new tensor with dimensions of [B, 1056, seq_len]. The tensor was transformed from 1D to 2D and sent to a 2D ResNet model. The 2D ResNet model is a deep convolutional neural network designed for processing 2D data, particularly with 2112 input channels. The architecture is characterized by a series of convolutional layers, residual blocks, and batch normalization. The initial convolutional layer, conv1, processes the input data with a 5x5 kernel, producing 64 output channels. Batch normalization (bn1) follows to normalize the output. The model's core consists of a sequence of residual blocks (layer0), each containing two convolutional layers with batch normalization. Notably, these residual blocks incorporate dilated convolutions with varying dilation rates (2, 4, 8, 16) and padding (2, 4, 8, 16, 32) in a hierarchical manner. The ResNet2D architecture consists of a total of 25 stacked 2D ResNet blocks, each incorporating two convolutional layers with batch normalization. The model concludes with a 2D dropout layer (out_dropout) with a dropout probability of 0.2 for regularization. The residual blocks is inspired by the ResNet architecture, allow for efficient training of deep networks by learning residual mappings. This ResNet2D architecture aims to capture features in 2D data. The use of dilated convolutions in the residual blocks enhances the model's receptive field, enabling the capture features in the input data. Additionally, the inclusion of dropout aids in preventing

overfitting during the training process, enhancing the generalization capability of the model.

After the 2D ResNet, the network branched into six separate paths. Each output branch consisted of a 2D convolution that projected down to dimensions of seq_len × seq_len × 37 (for 37 output bins), followed by a recurrent criss-cross attention (RCCA) module [35]. Within the RCCA module, attention queries and keys were projected to dimensions of seq_len × seq_len × 1 for each attention head. In our study, modal probability bin term refers to the bin within a probability distribution that has the highest predicted probability, representing the most likely value for each geometry-related output ($D_{C\alpha}$, ω, θ, and ϕ). For each residue pair, the network predicts a probability distribution across 37 bins, and we take the midpoint value of the modal probability bin to create real-valued matrices for these geometries. These real-valued distances and orientations are used to construct an initial backbone atom distance matrix (N, $C\alpha$, and C). The discretization of each output geometry involves 36 bins, with an additional bin specifically indicating distant residue pairs (dca > 18 Å). Predicted distances fall within the 0–18 Årange, with a bin width of 0.5 Å. Uniform discretization is applied to dihedral and planar angles, with bin widths of 10° and 5°, respectively. Each output branch includes a recurrent criss-cross attention module, facilitating the aggregation of information from all residue pairs within the output. Attention layers in protein structure prediction models offer a level of interpretability that is frequently absent in other aspects of the prediction process. The selection of criss-cross attention is motivated by its efficiency in aggregating information across a 2D grid, as exemplified by pairwise distance and orientation matrices. The criss-cross attention operation enables the network to attend to both rows and columns during the prediction phase for each residue pair. From the attention layer, the network predicted inter-residue distances between three pairs of atoms (C_α–C_α, C_β–C_β, N–O), as well as the set of inter-residue dihedrals (ω: C_α–C_β–C_β–C_α, θ: N–C_α–C_β–C_β) and planar angles (φ: C_α–C_β–C_β) [28].

Finally, the predicted values from the network were utilized in a fast Rosetta-based protocol for structure realization [34]. This protocol utilized the network's predictions to generate protein structures based on inter-residue distances and dihedrals, enabling the realization of protein structures using the network's predictions. Rosetta is a powerful software suite for computational biology, plays a crucial role in refining protein structures through energy minimization and enhancing the accuracy of 3D structure predictions. Employing sophisticated force fields, Rosetta models the potential energy of protein structures by considering factors such as bond angles, torsional rotations, and non-bonded interactions. Through Monte Carlo optimization techniques, Rosetta systematically explores the conformational space, allowing the protein to relax into states of lower potential energy. During training, we opted to use the focal loss [36] instead of the cross-entropy loss to improve the calibration of model predictions. Models trained with cross-entropy loss have been shown to overestimate the likelihood of their predicted labels [37]. We employed the Adam optimizer with a learning rate of 0.01, reducing the learning rate upon plateauing of validation loss to ensure stable training. To prevent overfitting, we employed the early stopping technique with a patience level of 15. The training was performed on an NVIDIA A100 80 GB PCIe GPU with 40 CPU cores and 754 GB RAM capacity. The batch size used for training was set to 8. All convolutions in the network are followed by ReLU activation. In total, the model contains about 9,209,788 trainable parameters. The model was trained on a nonredundant set of 2253 Fv structures from the Structural Antibody Database (SAbDab) [14,15] at 99% sequence identity. This dataset provides a diverse and representative set of antibody structures for effective training and evaluation. FvFold was implemented using PyTorch [38] and is freely available at github (https://github.com/pasangpythondiary/FvFold.git). Overall, our method employs a combination of 1D ResNet, Protein Transformer, and 2D ResNet models, along with focal loss optimization, early stopping, and extensive data preprocessing to predict inter-residue distances and angles crucial for determining the 3D structure of antibodies from their sequences.

4. Discussion

In this study, we have presented FvFold model, a novel method for predicting antibody structure. FvFold is based on protein language models, deep residual networks and Rosetta protocols. Our method is able to achieve highly accurate predictions for the antibody FV region. One of the main challenges we faced in developing FvFold was predicting the CDRH3 loop [10]. Predicting the structure of an antibody becomes challenging if the CDRH3 loop is either absent or excessively long [39].

In this study, we made decision to exclude certain PDB entries from our dataset due to issues such as missing structures or incomplete CDR loops. This approach was necessary to maintain the integrity and quality of the training data, ensuring that the dataset remained representative of high-quality antibodies. While this exclusion helps in developing a more accurate and reliable model, we acknowledge that it may result in an overestimation of the model's performance, as the evaluation does not fully account for the complexities present in some real-world antibodies. The ProtT5-XL-UniRef50 feature extraction process, combined with the subsequent deep learning computations, is inherently resource-intensive. Although parallel processing can significantly speed up predictions, it also leads to substantial GPU consumption. Our model also incorporates the Rosetta protocol for final structure prediction, which adds further computational complexity and dependencies. This level of resource demand may not be feasible for all users, particularly those with limited computational resources. It is important to note that the longer prediction times associated with our model are indicative of higher accuracy and the quality of the structures it predicts. The extensive computations involved in our approach contribute to better prediction performance, which is reflected in the detailed and accurate results generated by the model. For instance, the prediction times for FvFold are significantly longer, with an average runtime exceeding 15 min for paired antibody structure prediction. This contrasts with faster methods like IgFold, which can make predictions in around 23 s. While this substantial difference in runtime may limit the accessibility and practicality of FvFold, particularly for researchers with limited computational resources, it also underscores the trade-off between speed and accuracy in structural prediction methods. FvFold is also able to predict nanobody structures, even though it was not designed for this purpose and it was also able to outperform DeepAb model. The comparison has been shown in supplementary materials Table S1. The SAbDab-22H1-Nano benchmark was used to compare the nanobody performance which is also evaluated in tFold-Ab model [40]. This suggests that FvFold is a versatile method that could be adapted for other applications in the future.

5. Conclusion

FvFold demonstrates great capability in predicting antibody Fv 3D structure because we filtered and retained about 2200 PDB complete structures out of an initial pool of 7689 datasets, used only the experimental structure and ensuring the use of only the most reliable data sources for our predictive model. This emphasis on data quality undoubtedly contributed to the model's performance and efficacy. We constructed a deep learning model using protein transformer models, specifically the ProtT5-XL-UniRef50 model, which demonstrated superior predictive performance.

A key challenge with our 2D ResNet model is the risk of overfitting due to its complexity. To mitigate this, we used an ensemble approach by training the model five times independently and selecting the best weights from each run. This helps reduce overfitting by averaging out noise and capturing more generalized patterns. Additionally, we employed early stopping during training to halt when performance on the validation set no longer improved, further reducing overfitting. We also validated the model using different test and

validation sets to ensure it performed well on unseen data. Furthermore, testing on three standard benchmark sets assessed the model's robustness and generalizability across diverse scenarios, minimizing the risk of overfitting and enhancing its ability to generalize to new data. FvFold has been evaluated across three distinct benchmarks: the Therapeutic benchmark, the IgFold benchmark, and the Rosetta Antibody benchmark. Our results demonstrate that FvFold achieved the best performance for the CDRH3 loop in both the Therapeutic and IgFold benchmarks, showing a significant improvement compared to other methods. Specifically, the difference in performance on the Rosetta Antibody benchmark, where our model is slightly trailing by 0.08 Åin CDRH3 loop, is relatively minor and does not represent a substantial discrepancy. Moreover, the overall performance of FvFold across various loop types has been consistently strong, with superior results in other loop categories, concluding the competitive and reliable nature of our method.

The consistency of our results across these benchmarks demonstrates the model's accuracy and dependability, providing confidence in its potential for predicting antibody structures in the future. Moreover, our model could be really helpful for things like designing vaccines and supporting research centers. Because it can accurately predict antibody structures, it could be a valuable tool for future projects and studies.

Code availability

The code for this project is publicly accessible and can be found at https://github.com/pasangpythondiary/FvFold.git. It comes with comprehensive documentation for ease of understanding and utilization.

CRediT authorship contribution statement

Pasang Sherpa: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Kil To Chong:** Writing – review & editing, Validation, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Hilal Tayara:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The associated results and training data are available for public access and can be obtained from our dedicated repository at https://zenodo.org/records/10791148. The repository comprises several files.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C2005612) and (No. 2022R1G1A1004613) and in part by the Korea Big Data Station (K-BDS) with computing resources including technical support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109128.

References

[1] P. Nelson, G. Reynolds, E. Waldron, E. Ward, K. Giannopoulos, P. Murray, Demystified…: monoclonal antibodies, Mol. Pathol. 53 (3) (2000) 111.
[2] R.-M. Lu, Y.-C. Hwang, I.-J. Liu, C.-C. Lee, H.-Z. Tsai, H.-J. Li, H.-C. Wu, Development of therapeutic antibodies for the treatment of diseases, J. Biomed. Sci. 27 (2020) 1–30.
[3] Z. Wang, G. Wang, H. Lu, H. Li, M. Tang, A. Tong, Development of therapeutic antibodies for the treatment of diseases, Mol. Biomed. 3 (1) (2022) 35.
[4] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al., Protein complex prediction with AlphaFold-multimer, Biorxiv (2021) 2021–2010.
[5] J.A. Ruffolo, J.J. Gray, Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies, Biophys. J. 121 (3) (2022) 155a–156a.
[6] J.A. Ruffolo, J. Sulam, J.J. Gray, Antibody structure prediction using interpretable deep learning, Patterns 3 (2) (2022).
[7] B. Abanades, G. Georges, A. Bujotzek, C.M. Deane, ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation, Bioinformatics 38 (7) (2022) 1877–1880.
[8] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, Adv. Neural Inf. Process. Syst. 34 (2021) 29287–29303.
[9] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al., Prottrans: Toward understanding the language of life through self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (10) (2021) 7112–7127.
[10] B.D. Weitzner, J.J. Gray, Accurate structure prediction of CDR H3 loops enabled by a novel structure-based C-terminal constraint, J. Immunol. 198 (1) (2017) 505–515.
[11] J.A. Ruffolo, C. Guerra, S.P. Mahajan, J. Sulam, J.J. Gray, Geometric potentials from deep learning improve prediction of CDR H3 loop structures, Bioinformatics 36 (Supplement_1) (2020) i268–i275.
[12] N.A. Marze, S. Lyskov, J.J. Gray, Improved prediction of antibody VL–VH orientation, Protein Eng. Des. Sel. 29 (10) (2016) 409–418.
[13] M.I. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A.P. Lewis, A. Bujotzek, J. Shi, C.M. Deane, Five computational developability guidelines for therapeutic antibody profiling, Proc. Natl. Acad. Sci. 116 (10) (2019) 4025–4030.
[14] M.I. Raybould, C. Marks, A.P. Lewis, J. Shi, A. Bujotzek, B. Taddese, C.M. Deane, Thera-SAbDab: the therapeutic structural antibody database, Nucleic Acids Res. 48 (D1) (2020) D383–D388.
[15] J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, C.M. Deane, SAbDab: the structural antibody database, Nucleic Acids Res. 42 (D1) (2014) D1140–D1146.
[16] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (1) (2000) 235–242.
[17] B.D. Weitzner, J.R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R.L. Dunbrack, J.J. Gray, Modeling and docking of antibody structures with Rosetta, Nat. Protoc. 12 (2) (2017) 401–416.
[18] C. Chothia, A.M. Lesk, Canonical structures for the hypervariable regions of immunoglobulins, J. Mol. Biol. 196 (4) (1987) 901–917.
[19] G. Georgiou, G.C. Ippolito, J. Beausang, C.E. Busse, H. Wardemann, S.R. Quake, The promise and challenge of high-throughput sequencing of the antibody repertoire, Nature Biotechnol. 32 (2) (2014) 158–168.
[20] J. Dunbar, K. Krawczyk, J. Leem, C. Marks, J. Nowak, C. Regep, G. Georges, S. Kelm, B. Popovic, C.M. Deane, SAbPred: a structure-based antibody prediction server, Nucleic Acids Res. 44 (W1) (2016) W474–W478.
[21] J.C. Martins, S. Ballet, et al., Downsizing antibodies: Towards complementarity-determining region (CDR)-based peptide mimetics, Bioorg. Chem. 119 (2022) 105563.
[22] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, A. Rives, Transformer protein language models are unsupervised structure learners, Biorxiv (2020) 2020–2012.
[23] S. Targ, D. Almeida, K. Lyman, Resnet in resnet: Generalizing residual architectures, 2016, arXiv preprint arXiv:1603.08029.
[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[25] N. Deshpande, K.J. Addess, W.F. Bluhm, J.C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, L. Xie, L. Chen, Z. Feng, et al., The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema, Nucleic Acids Res. 33 (suppl_1) (2005) D233–D237.
[26] W. Gao, S.P. Mahajan, J. Sulam, J.J. Gray, Deep learning in protein structural modeling and design, Patterns 1 (9) (2020).
[27] S. Bhattarai, K.-S. Kim, H. Tayara, K.T. Chong, Acp-ada: a boosting method with data augmentation for improved prediction of anticancer peptides, Int. J. Mol. Sci. 23 (20) (2022) 12194.
[28] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations, Proc. Natl. Acad. Sci. 117 (3) (2020) 1496–1503.

[29] J. Xu, M. Mcpartlon, J. Li, Improved protein structure prediction by deep learning irrespective of co-evolution information, Nat. Mach. Intell. 3 (7) (2021) 601–609.
[30] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (7792) (2020) 706–710.
[31] R.W. Floyd, Algorithm 97: shortest path, Commun. ACM 5 (6) (1962) 345.
[32] J. Zhang, Q. Wang, B. Barz, Z. He, I. Kosztin, Y. Shang, D. Xu, MUFOLD: A new solution for protein 3D structure prediction, Proteins Struct. Funct. Bioinform. 78 (5) (2010) 1137–1152.
[33] J.K. Leman, B.D. Weitzner, S.M. Lewis, J. Adolf-Bryfogle, N. Alam, R.F. Alford, M. Aprahamian, D. Baker, K.A. Barlow, P. Barth, et al., Macromolecular modeling and design in Rosetta: recent methods and frameworks, Nature Methods 17 (7) (2020) 665–680.
[34] A. Leaver-Fay, M. Tyka, S.M. Lewis, O.F. Lange, J. Thompson, R. Jacak, K.W. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, et al., ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, in: Methods in Enzymology, vol. 487, Elsevier, 2011, pp. 545–574.
[35] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[37] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. Torr, P. Dokania, Calibrating deep neural networks using focal loss, Adv. Neural Inf. Process. Syst. 33 (2020) 15288–15299.
[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[39] M.L. Fernández-Quintero, J. Kraml, G. Georges, K.R. Liedl, CDR-H3 loop ensemble in solution–conformational selection upon antibody binding, in: MAbs, vol. 11, (6) Taylor & Francis, 2019, pp. 1077–1088.
[40] J. Wu, F. Wu, B. Jiang, W. Liu, P. Zhao, tFold-Ab: fast and accurate antibody structure prediction without sequence homologs, Biorxiv (2022) 2022–2011.